修 士 論 文 の 和 文 要 旨

研究科・専攻	大学院　情報理工学研究科　総合情報学専攻　博士前期課程		
氏　　名	若山　裕介	学籍番号	1230079
論 文 題 目	接続詞の必要箇所の自動判定 ―文章添削システムの構築に向けて―		

要　　旨

文章を作成する能力というのは正確に物事を伝える上でとても重要である．その際，自然な文章にするためには適切な接続詞が必要である．我々は文章の自動添削システムを構築することを最終的な目標とし，その中でも特に判定が困難である，「論理関係の適切さ」の判別を行う．そこで，様々な観点からどのような場合に接続詞が必要であるかを分析することで，任意の文書を与えた時に接続詞を必要とする箇所を自動的に判定する手法が必要であり，本研究ではその手法の開発を目的とする．

本研究では機械学習の分類器を用いて，接続詞がどのような文間で必要になるかを自動的に推定する．接続詞が必要となる箇所は必須箇所と推奨箇所の 2 つに区別することができると考える．そのため必須，推奨箇所について独立に分類器を構築する．使用した属性は助詞（9 種類），助動詞（9 種類），繰り返し語句（名詞，動詞），シソーラス距離，文節パタン情報，係り受けのパタン情報である．

必須箇所では，決定木では適合率が高い結果が得られた．そのため，分類器が接続詞が必要である箇所と判断するものに対しては正しい結果を得られると言える．一方，SVM では決定木よりは再現率が高いものの，適合率は低くなった．そのため，添削システムとして用いるのであれば SVM よりも決定木を用いる方が有効である事がわかった．

推奨箇所では，決定木では必須箇所同様に「省略できない」と分類されたものに関しては正しく分類出来た．これは，適合率がベースラインよりも高い結果から有効である事がわかった．

文章添削システムに向けて，既存の研究である接続関係の同定と本研究の分類器を適用する前後での分類性能の結果を出した．この結果を見ると「並列」，「累加」，「転換」，「逆接」に関しては分類性能が上昇したため，本研究が有効に働いたと考えられる．

平成25年度総合情報学専攻修士研究論文

接続詞の必要箇所の自動判定
–文章添削システムの構築に向けて–

提出年月日： 平成26年01月30日

提出者： 学籍番号 1230079

若山 裕介

指導教員： 内海 彰 教授

尾内 理紀夫 教授

目 次

1　はじめに

文章を作成する能力というのは正確に物事を伝える上でとても重要である．そこで，計算機が文章を自動的に添削するシステムが必要となる．

計算機が文章を添削するときに重要なこととして，以下の点があげられる．

- 文章表現の適切さの判別
  - 文の難易度（リーダビリティ）
  - 文の長さ
- 論理展開の適切さの判別
  - 接続関係の判別（例：逆説，累加等）
  - 起承転結の判別
- 文章の構造的な部分
  - 事実と意見の判別
  - 客観表現のための具体的数値の判別
  - 意味として一意かどうかの判別

この一覧のなかで，文の難易度を測る研究は盛んに行われている [2, 3]．また，文の長さ等を利用して日本語の難易度を公式として求めた研究 [4] も存在する．しかし，これらは文の読みやすさ等に注目しているのみであり，添削をする上で指摘することとしては不十分である．

文章を自然な形にするための重要なタスクとして論理展開の適切さの判別があげられる．それらの研究の一つに小論文を自動的に添削するシステムが存在する [5]．ここでは，小論文に特化することで論理展開として接続詞等を手掛かり語として用い判定することや，LSA(Latent Semantic Analysis) 等を用い問題文と解答文との類似度を測ることで実現している．

しかし，以上の研究等で手掛かり語として使用している接続詞は正しく利用されているという前提に成り立っている．

この問題の例として，近年大学生等でも接続詞を上手に扱えていない人が増えている．そこで，計算機が文章中の全ての文間において接続詞が必要である箇所を推定し，接続詞が必要な箇所では結束関係を正しく表す接続詞を自動的に判断して，可能ならば正しい接続詞を指摘する添削システムが必要となる．

言語処理において接続詞は文間の結束関係を同定するための手掛かり語として使用されることが主であり，接続詞自体を同定する研究は少ない．しかし，先述した添削システムを実現するには接続詞を同定することが必要であるが，同じような関係を表す接続詞どうしの違いが曖昧であることから計算機で同定するのはとても困難である．そのため，接続詞を大まかに分類した接続関係を同定することが求められる．そこで我々は文間の接続詞（を分類した接続関係）の同定に関する研究 [1] を行った．しかし，この研究では文間に接続詞が入るという前提で行ったため，文間に接続詞が必要ない場合における対処が問題となる．接続詞がない文の同定を,「接続詞なし」というカテゴリーを追加して学習することも考えられるが，接続詞がない文間でも接続関係は存在する．また実際に接続詞が出現する頻度は少ない（10 %程度）ために，接続詞が必要ないことの同定は非常に困難である．そのため，接続詞が必要であるかどうかは接続関係の同定とは全く別問題であると言える．そこで，様々な観点からどのような場合に接続詞が必要であるかを分析し，任意の文書を与えた時に接続詞を必要とする箇所を自動的に判定する手法が必要であり，本研究ではその手法の開発を目的とする．

先述したように，接続関係を同定する研究 [1, 7] は少なく，ましてや接続詞自体が省略できるかどうかの研究は行われていない．しかし，言語学や国語教育などの分野においては接続詞の省略について分析がいくつか行われている．市川 [9] はどのような接続詞が省略しやすいかを,「接続詞を挟む 2 文」と「ある文書内での接続詞」に対して調査を行っている．この調査では，文の構造や文脈の有無などに応じて接続詞が省略可能かどうかが決定されることが明らかになっている．

よって，本研究ではこれらの知見を考慮して，大量の実験データから接続詞が省略されやすい状況を分析し，接続詞が必要であるかを判定する手法を開発・提案する．

2　関連研究

ここでは，関連研究について述べる．本研究の目的である接続詞が必要である箇所を推定する研究に直結した関連を 2.1 節に，本研究とは直接関係ないが接続詞が必要であると判定されたものに対して，次のステップとしてどのような接続詞（又はそれを分類した接続関係）を同定させるための関連研究についてを 2.2 に述べる．

2.1　接続詞の必要箇所について

言語処理においては，接続詞の必要箇所を研究したものは存在しない．しかし，国語学や教育学においては研究 [9] がされている．ここでは，このことについて詳しく説明する．

市川 [9] は文と文をつなぐ形式として，次のような種類があると述べている．

1. 前後の文（あるいは節）相互を直接，論理的に関係付ける形式
   次の 4 形式は例えば「順接」とか「逆接」などのように，文と文や節と節との論理的関係を様々に示し，前後を接続する．本研究においてはここを対象としている．これらの形式において用いられる語句は，品詞上で異なっていても，文脈展開の上では，同様な機能をもつ．

   (a) 接続詞を用いる
   例文：その計画はなかなかおもしろい。しかし、実行は困難だと思う。

   (b) 接続詞的機能をもつ語句を用いる（接続語句）

      i. 接続して気に用いられる副詞・名詞
      例文：そんなことは子どもでも知っている。いわんや、おとなたちの知らぬはずはない。

      ii. 接続詞的に用いられる連語
      例文：朝から雨がはげしかった。そのため、競技会は延期された。

   (c) 接続助詞を用いる
   例文：山には雪が降ったが、里には降らなかった。

(d) 接続助詞的機能をもつ語句を用いる
例文：力が足りなかった ため、不成功に終わった。

2. 前文（あるいは前節）の内容を後文（後節）の中に持ち込んで、前後を内容的に関係付ける形式
次の 3 形式は，前後の論理的関係を示すのではなく，前文（節）の内容を，後文（節）の内容として持ち込んで，内容の上から両者を関係づけるところに，その特徴が見られる．

(a) 指示語を用いる
例文：かれは駅のベンチで本を読んでいた。それ を見たものがあった。

(b) 前文の語句と同一の語句を用いる
例文：窓からは林が見えた。林 は、夕日に美しく照らされていた。

(c) 前文の語句に対して同義あるいは類義の語句を用いる
例文：山の上には最前から巨大な入道雲がそびえている。雲の峰 は、少しずつ形を変えはじめた。

3. その他の形式
次の 5 形式は，前後の文（節）のつながりを説明ないし暗示するものと考えられ，前述の中間的な性質をもつものである．

(a) 前後関係を説明する表現を用いる
1(b)ii では，全体として関係的意味を表し，形式上肯定化しているのに対して，ここでは多分に実質的意味を表し形式上もかなり自由でいろいろな言い方ができる．

(b) 前文の表現を（要約して）接続語的に繰り返す
例文：朝から雨が降った。雨によって、大会は延期された。（順接的関係）

(c) 特殊な文末表現を用いる
例文：わたしは返答に困った。思いもよらないことだった からである。（理由の補足）

(d) なんらかの意味で前後関係を表す後（もしくは記号）を用いる

i. ある種の助詞
例文：雨が激しくなった。風 さえ 加わってきた。（追加）

ii. ある種の名詞
例文：以上が前半の内容のあらましです。次 は、後半についてです。（序列）

iii. 「付記」「付（つけたり）」「二伸」「イ・ロ・ハ」「1・2・3」「a・b・c」などの表示

(e) 特殊な活用形を用いる。

i. 連用形中止法
例文：朝五時に 起き、すぐ支度にとりかかった。

ii. 仮定形
例文：行くのが いやなら、よしなさい。（仮定条件にもとづく順接）

このなかでも 1 節における接続詞や接続詞的機能をもつ語句では，表現する際用いたり用いなかったりする．推敲する場合にも，接続語句を加えることもあれば，削除することもある．接続語句を省略する場合でも，どれでも一様に省略しうるというものではない．

そこで，どのようなときに接続語句が省略しやすいのかを過去の知見と客観的実験を行なっている．佐久間鼎氏は,「接続詞の省略」について次のように述べている．事態が順当に進行して，前後が適合の関係にある場合や，事態が自然に納得されるとか，他の方法によって直知されるとかいう場合には，接続詞は省略されうるが，前後が背馳するような場合や論理的な関係を明示する必要のある場合には，省略するわけにはいかないと述べている．

次に客観的な実験では二種類の調査を行なっている．調査 A では，各種の文脈モデルについて，調査 B では実際の文脈例について，それぞれの接続語句を省略しうるか否かを調査している．文の続き方がわからくならない限度内で，省略できるかどうかの調査である．被験者は大学生であり調査 A では 65 名，調査 B では 53 名用意している．

調査 A では 23 文用意し全ての文章で文間に入る接続詞は異なっている．表 1 に接続詞を省略しても良いと答えた数の大きい順を示す．次に調査 B では 2 つの少し長

表 1: 調査 A の接続詞が省略しやすいもの一覧

接続詞	割合	接続詞	割合	接続詞	割合
また	100%	一方	95%	すなわち	92%
それとも	82%	なお	81%	まして	75%
そして	71%	かくて	64%	つぎに	60%
すると	58%	要するに	58%	ところで	43%
ただし	40%	なぜなら	40%	それなのに	25%
さて	22%	とりわけ	22%	そのうえ	14%
そのため	14%	しかし	13%	ともあれ	6%
ところが	5%				

表 2: 調査 B の接続詞が省略しやすいもの一覧

接続詞	割合	接続詞	割合	接続詞	割合
つまり	96%	そうして	94%	ところが	72%
しかし	70%	やがて	60%	もっとも	58%
ところで	55%	さらに	53%	こうして	47%
だから	34%	そのために	13%		

い文章を読ませ，その中に入る接続詞について省略しても良いかどうかの調査を行なっている．表 2 に接続詞を省略しても良いと答えた数の大きい順に示す． 接続語句を省略する場合は主観的要素が伴ってしまうが，多数の被験者を用意することで客観的な傾向が得られる．

結果を見ると省略率が高いものと低いものが存在することが分かる．これは接続語句の省略率が高いのは文間に接続語句がなくても，比較的容易に二文のつながりが把握されるという事である．この場合は，接続語句以外の，二文の関係を暗示するような語句が用いられている場合や内容相互のつながりが自然にすぐわかるなど，接続語句を必ずしも必要としないからである．

このように，接続語句ごとに補助的用法をもつものと必須的用法を持つものに分類できる．しかし，それらを各接続語句の意味では分類することはできない．これは，文脈を考慮した場合や 2 文における他の手がかりが存在しなければ省略するこ

とが出来ないからである.

まとめとして接続語句はその用法として補助的な場合と必須的な場合とが区別され，接続語句によってそのどちらの傾向が強いかが一応指摘できるが，それは文脈によってかなり浮動し必須的なものが補助的なものに，補助的なものが必須的なものに転化するという場合があると述べている.

2.2 接続関係について

ここでは先述したSVMを用いた手法[1]と用例利用型の手法[6]と英語圏における研究であるEmily[7]について詳しく説明する．

2.2.1 文間の接続関係

先行研究であるSVMを用いた手法[1]と用例利用型[6]では，日本語の接続詞の分類について言及している．日本語の接続関係は文献[9]から大きく8種類（「順接」「逆説」「添加」「対比」「転換」「同列」「補足」「連鎖」）に分類している．この中で以下の理由から最終的に6種類（「並列」「例示」「因果」「累加」「転換」「逆説」）に分類している．

1. 「同列」と「補足」,「逆説」と「対比」を区別するのは容易ではなかったため一緒にした．
2. 「例示」は市川の類型にはないが，分自体に特徴があったため，別に分類したほうが良いと考えた．
3. 「連鎖」は市川の分類で定義だけはあるものの，具体的には触れられていないとして削除した．

接続詞を分類した表を3に示す．ここでは，代表的な接続詞だけを挙げた．また，この分類に入らない接続も少なからず存在する．例えば,「おしむらくは」や「なかんずく」などである．しかし，これらの接続詞は他の接続詞に比べて新聞コーパスなどで出現する頻度がとても少ないことから，文章を書く上で使用する頻度が少ないと考えた．このような考えから，本研究ではこれらの接続詞は対象外とした．

また，多義性のある接続詞も存在する．これは表3に示した種類の複数に属することを意味する．しかし，本研究の目的は一意に決めることであることや多義性のある接続詞は全体の約1割程度である．これらの場合は接続関係全てに属することにしている．

表 3: 接続詞の分類

種類	接続詞の例
並列	一方，もしくは，あるいは，つまり，…
例示	例えば
因果	だから，ゆえに，なので，すると，…
累加	また，そして，さらに，しかも，まずは，…
転換	さて，ところで，では，…
逆説	しかし，だが，でも，なのに，ところが，…

表 4: 特徴ベクトルに用いた素性

種類		重み	次元数
品詞	名詞-サ変接続	文中に現れる頻度	8865
	動詞	文中に現れる頻度	7600
	助詞	文中に現れる頻度	313
	形容詞	文中に現れる頻度	958
	接続詞 (判定する接続詞は除く)	文中に現れる頻度	208
文間類似度		類似度 (実数)	1
機能表現の意味情報		1	94
係り受け解析のパタン情報		1	551888
シソーラスを用いた名詞-サ変接続の種類分け		文中に現れる頻度	8865

2.2.2 SVM を用いた手法

まず，SVM の用いた手法 [1] の研究で対象としている文書は新聞記事である．概要を説明する．まず，接続詞を挟む前後各数文（この数は変更可能）を情報として入手する．これらの文すべての集合を 1 単位とする．各文に形態素解析・係り受け解析・パターンマッチ等の各処理を行い，表 4 に示した素性に基づき 1 単位ごとに特徴ベクトルを構成する．なお形態素解析には MeCab[1]，係り受け解析には CaboCha[2]を用いる．

[1]http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[2]http://code.google.com/p/cabocha/

与える情報として，接続詞を挟む前後の文数と素性ごとに前後の区別も付与できるようになっている．例えば，前後 2 文を情報として用いる時，前一文目と前二文目をそれぞれ区別する場合，前文と後文の 2 つに分ける場合，さらに前文と後文の区別もしないという 3 通りが考えられる．

以下に表 4 の素性の詳細を述べる．品詞の情報としては，名詞，動詞，助詞，形容詞，接続詞（判定しようとしている接続詞は除く）を用いた．また，名詞に関してはシソーラス[3] を用いて，名詞をある階層に分類することも考える．

文間類似度は接続詞を挟む前後の文に含まれる単語の $tf \cdot idf$ によるベクトル（素性による特徴ベクトルとは関係ない）の類似度のことである．文ベクトルを d_i, d_j とすると，類似度 $S(d_i, d_j)$ は以下のようになる．

$$d_i = (\omega_{i1}, \omega_{i2}, \cdots, \omega_{iM})^T, \omega_{ij} = tf_{ij} \cdot idf_j \tag{1}$$

$$tf_{ij} = \text{文 } i \text{ 中に出現する単語 } j \text{ の頻度} \tag{2}$$

$$idf_j = \log \frac{\text{総文数}}{\text{単語 } j \text{ が出現する文数}} + 1 \tag{3}$$

$$S(d_i, d_j) = \frac{\Sigma \omega_{1j} \omega_{2j}}{\sqrt{\Sigma \omega_{1j}^2} \sqrt{\Sigma \omega_{2j}^2}} \tag{4}$$

「機能表現の意味情報」とは，機能表現辞書「つつじ」[11] に存在する,「立場」「許可」などの意味情報のことである．これらの意味情報は全部で 94 種類ありそれらに該当する機能表現が文に存在した場合，該当する意味情報の素性が 1 となる．

係り受けの構文パタンとは，文を係り受け解析しその文のつながりをパタン化したものである．具体的には以下の手順で構文パタンを求める．

1. 文を係り受け解析し，各文節ごとに文節末が助詞，助動詞，すべての品詞で「非自立」であるものと動詞の「ある」についてのみ抽出する．その他の品詞についてはワイルドカードにする．
2. 上記で求めた文節パタンのすべての係り受けの組み合わせを求め，それらを構文パタンとする．

これらにより作成されたベクトルを用いて 6 種類の接続関係に分類する分類器を SVM により作成する（多値分類手法は One vs Rest 法）．組み合わせた方法は表 5 で

[3] 日本語語彙大系，岩波書店 (1997)

あり，それらの結果は表 6 のようになる．これらの結果を見ると全体として 45%程度は正しく分類できていることがわかる．

表 5: 組み合わせた素性一覧（○：前後で完全に区別する，△：前後で大別する，□：前後で区別しない，空欄：用いない）

種類	扱う文数	品詞					文間類似度	機能表現の意味情報	係り受けの構文パタン
		名詞-サ変	動詞	助詞	形容詞	接続詞			
素性集合 1	前後各 2 文	△	△	△	△	△	△	□	○
素性集合 2	前 1 文後 2 文	△	△	△	△	△	△	□	○
素性集合 3	前 1 文後 2 文	△	△	△	△	△	△	□	△
素性集合 4	前後各 1 文	△	△				△	□	
素性集合 5	前後各 2 文	△	△				△	□	

表 6: SVM を用いた手法 [1] と用例利用型 [6] での結果（F 値）

手法	並列	例示	因果	累加	転換	逆説	平均
提案手法 素性集合 1	**0.529**	0.368	0.356	0.402	**0.509**	**0.447**	0.437
提案手法 素性集合 2	0.505	**0.466**	0.357	0.304	0.437	0.420	0.418
提案手法 素性集合 3	0.371	0.398	**0.447**	0.384	0.447	0.410	0.413
提案手法 素性集合 4	0.212	0.193	0.296	0.188	0.266	0.204	0.230
提案手法 素性集合 5	0.414	0.371	0.204	0.262	0.269	0.277	0.300
提案手法 素性集合 1+2	**0.529**	0.453	0.376	**0.416**	0.500	0.441	**0.452**
提案手法 素性集合 1+2+3	0.527	0.420	0.371	0.386	0.440	0.413	0.426
用例利用型 一意	0.517	0.337	0.300	0.409	0.205	0.418	0.365
用例利用型 ランダム	0.350	0.230	0.241	0.273	0.232	0.267	0.266

注）太字は各接続関係ごとに精度が一番良いものを表す.

2.2.3 用例利用型

まず，用例利用型の手法 [6] の研究で対象としている文書は Web 文書である．概要を説明する．

1. 接続詞を挟む，前後各 1 文のみについて考える
2. まず，前後 2 文の述語と述語にかかる格要素 $V_{ij}, N_{ij}(i = 1, 2)$(j は文書番号) を抽出する
3. 係り受けのパタン情報を前後 2 文でそれぞれ求め，それらのすべての組み合わせを構文パタンとする．
   ここで，構文パタンの要素数の合計をパタンスコアと呼ぶ
4. 次に前後各 1 文の述語と述語にかかる格要素 V_{ij}, N_{ij} を後述する素性に基づいてそれぞれ 4 種類のクラスタリングを行う
5. クラスタリングされた結果を素に入力文 i と候補文 c とのスコア $S_{V1}(i, c), S_{V2}(i, c), S_{N1}(i, c), S_{N2}(i, c)$ をそれぞれ計算する
6. 計算したスコアのうち最も高い候補文の接続関係をシステムの解とする
   ここで，スコアが同一の値が出た場合は複数の回答を認めている（この研究では複数回答が多かったという結果が出ている）

次に，クラスタリングについて述べる．

まず，素性は述語 V_{ij} と格要素 N_{ij} で異なっている．それぞれ表 7,8 に示す．ここで，target と source という概念が存在する．例えば，1 文目の述語 V_{ij} をクラスタリングする場合，1 文目のことを target，2 文目のことを source と呼ぶ．逆に 2 文目をクラスタリングする場合は，2 文目が target，1 分目が source となる．

この方法で用意した 10000 文のコーパスからクラスタリングを行う．ここでのクラスタリングは階層的ベイズクラスタリングを用いる．階層的ベイズクラスタリングとは，多次元のデータセットに対して，要素間の類似度に基づいて比較的「近い」要素群をクラスタとして発見する分析手法の一つである．ここでは，各アイテム（述語と述語にかかる格要素）の中で最も「近い」ものから順にボトムアップで各アイ

表 7: 述語のクラスタリングに用いた素性と重み

重み	素性
2	target の述語
5	source の述語
2	source の述語にかかるすべての形態素（記号を除く）
10	文間の接続詞
1	係り受け解析のパタン情報のパタンスコアが 4 以上のもの

表 8: 述語にかかる格要素のクラスタリングに用いた素性と重み

重み	素性
2	target の述語にかかる格要素
5	source の述語
2	source の述語にかかる格要素
2	source の述語にかかるすべての形態素（記号を除く）
10	文間の接続詞
1	係り受け解析のパタン情報のパタンスコアが 4 以上のもの

テムをまとめ上げていく．一般的な木構造とは違い，同じ階層は存在せず，指定したクラスタ数になるように分割点を決定する．

クラスタリングの分割停止条件としてエントロピーを用いている．エントロピーは以下の式により求められる．

$$H(D) = -\sum_{\text{接続関係 } i} P_i \log_2 P_i \tag{5}$$

また，分割後のエントロピーは各クラスタ中のエントロピーの加重平均で表される．例えば分割前のデータを D_0，分割後のデータを D_1, D_2 とすると

$$H(D_1 + D_2) = \frac{|D_1|}{|D_0|} H(D_1) + \frac{|D_1|}{|D_0|} H(D_2) \tag{6}$$

と表せられる．さらに以下の条件を加えている．

1. 任意の単語 A がすべてのクラスに属す場合
2. 1 種類の単語だけで構成されているクラスが存在する場合

この条件を満たすと，分割停止となる．

最後にスコアの計算方法を説明する．

述語に関するスコア S_{Vi} については以下のように求める．

入力文 i，候補文を c としたときのスコア $S_{V1}(i,c)$ の初期値を 0.001 とする．この初期値から以下の図 1 のアルゴリズムにより求める．次に述語にかかる格要素に関するスコア S_{Ni} について説明する．

述語にかかる格要素のスコア S_{Ni} も初期値を 0.001 として以下の図 2 のアルゴリズムにより求める．

以上のように算出したスコアを以下の式で最終的なスコアとする．

$$Sim(i,c) = (S_{V1}(i,c) \times S_{V2}(i,c)) \times (S_{N1}(i,c) \times S_{V1}(i,c)) \times (S_{N2}(i,c) \times S_{V2}(i,c)) \quad (7)$$

```
if( "V1c" = "V1i" ){
   SVI(i ,c ) += 1;
}else{
   if( "V1cとV1iを含むクラスタが存在する" ){
      SVI(i ,c ) += 1/( V1 cとV1iを含むクラスタの数 );
   }
}
```

図 1: 述語による単語スコアの加算方法

```
if( "N1c" = "N1i" ){
   SNI(i ,c ) += 1;
}else{
   if( "N1cとN1iを含むクラスタが存在する" ){
      SNI(i ,c ) += 1/( N1cとN1iを含むクラスタ数 );
   }else{
      if("N'1c" = "N'1i"){
         SNI(i ,c ) += 0.1;
      }else{
         if( "N'1cとN'1iを含むクラスタがある" ){
            SNI(i ,c ) += 0.1/( N'1 cとN'1iを含むクラスタの数 );
         }
      }
   }
}
```

図 2: 述語にかかる格要素による単語スコアの加算方法

2.2.4 英語圏における研究

英語圏の研究には，Pitler ら [7] の研究が存在する．ここでは，but や because 等のあからさまな接続詞ではなく，それら以外の情報で結束関係に影響がある素性について調べている．結束関係の種類は,「Comparison（比較）, Contingency（随伴）, Expansion（展開）, Temporal（時間)」である．ここで，随伴とはキーワードの共起に注目することを言う．結束関係を導く手段として，ナイーブベイズ分類器を用いている．その際素性としては以下の項目を挙げている．また，見ている文情報はある文とある文の結束関係を見ているため 2 文である．

1. 極性判定 (Polality Tags)
2. 質問形判定 (Inquirer Tags)
3. お金/パーセント/数字 (Money/parsent/Num)
4. WSJ-LM
5. Expl-LM
6. 動詞
7. First-Last, First3
8. モダリティ
9. 文脈
10. 単語のペア

これらをそれぞれの素性のみを用いて分類させ，出た結果からその結束関係で有効でありそうな素性をみつけるという研究である．結果を表 9 に載せた．

表 9: f-score(accuracy) of various features sets; Naive Bays.[7]

Comp vs. Other	
Wordpairs-TextRels	17.13 (46.62)
Wordpairs-PDTB Expl	19.39 (51.41)
Wordpairs-PDTB Impl	20.96 (42.55)
First-last, first3 (*best-non-wp*	21.01 (52.59)
Best-non-wp L Wordpairs-selected	21.88 (56.40)
Wordpairs-selected	21.96 (56.59)
Cont vs. Other	
Wprdpairs-TextRels	31.10 (41.83)
Wordpairs-PDTB Expl	37.77 (56.73)
Wordpairs-PDTB Impl	43.79 (61.92)
Polarity, verbs, first-last, first3, modality, context (*best-non-wp*)	42.14 (66.64)
Wordpairs-selected	45.60 (67.10)
Best-non-wp + Wordpairs-selected	47.13 (67.30)
Expn vs. Other	
Best-non-wp + wordpairs	62.39 (59.55)
Wordpairs-PDTB Impl	63.84 (60.28)
Polarity, inquirer tags, context(best-non-wp)	76.42 (63.62)
Temp vs.Other	
First-last,first3 (*best-non-wp*)	15.93 (61.20)
Wordpairs-PDTB Impl	16.21 (61.98)
Best-non-wp + Wordpairs-PDTB Impl	16.76 (63.49)

3　提案手法

機械学習の分類器を用いて，接続詞がどのような時に必要になるかを自動的に推定する．ここで，接続詞の必要の度合いを以下のように考えた．

1. 必要
2. ある方が良い
3. あっても良い
4. なくても良い
5. ない方が良い
6. 絶対ない

我々は接続詞が必要となる文間は必須箇所と推奨箇所の 2 つに区別することができると考えた．

必須箇所とは文間にある接続詞がなければ接続関係が理解できなくなるほど文の構成として重要な文間である．（上記 1 と 2 の境界）

推奨箇所は必須箇所の条件に加えて文としてのわかりやすさを考慮した上で接続詞を入れる文間であると考えた．（上記 3 と 4 の境界）

以上のような違いを考慮し，それぞれの箇所について独立に分類器を構築する．

3.1 分類に用いる属性

ここでは，機械学習による分類器を構築するため，どのような属性を学習させるかについて述べる．以下に，本研究で提案する属性について述べる．

- 終助詞や格助詞など 9 種類に分類した助詞：前文と後文のそれぞれにおける 9 種類の助詞の出現頻度
- 断定や受身・使役など 9 種類に分類した助動詞：前文と後文のそれぞれにおける 9 種類の助動詞の出現頻度
- 繰り返し語句（名詞）：前文で出現した名詞が後文で出現した延べ回数
- 繰り返し語句（動詞）：前文で出現した動詞が後文で出現した延べ回数
- 前文の主語と後文の主語に含まれる名詞間のシソーラス距離の最小値
- 前文に含まれる名詞と後文に含まれる名詞のシソーラス距離の最小値
- 文節パタン情報
- 文節数が 2 から 4 までの係り受けパタン情報

3.1.1 助詞，助動詞

分類に用いる属性を決定する上で文献 [9] において人が文章を読むときに文の前後関係をつかみやすくする要素として挙げられていた助詞や助動詞の属性を用いた．助詞や助動詞をそのまま表層的に学習することも考えられるが，これらを表 10 のように分類したものを用いる．なお，助詞，助動詞の分類は文献 [9, 10] を参考にした．

3.1.2 シソーラス距離

名詞を体系的にまとめた日本語語彙体系（シソーラス）[4] を用いて名詞の距離を測る．これは文間にある接続関係を推測できるような単語は接続詞の必要に寄与していると考えられるからである．例えば，以下のような場合である．

[4] 日本語語彙体系，岩波書店 (1997)

表 10: 助詞，助動詞の分類

品詞	種類
助詞	副助詞，終助詞，格助詞，係助詞，接続助詞， 並立助，連体化，副詞化，特殊
助動詞	断定，受身・使役，否定，完了・過去， 推量，丁寧，希望，比況，その他

- 兄は右の道を進んだ。一方、弟は左の道を選んだ。

例文のアンダーラインの部分は「右」と「左」という対象的な語彙があるため接続詞「一方」は必ずしも必要ではないと言える．ここでシソーラスは名詞が全 12 階層にわかれており，木構造になっている．この例では，第 6 階層目に「左右」というノードがあるため，この 2 つのシソーラス距離は 0 となる.（同じノードにあるため）

このようなことから，本研究では「前文の主語と後文の主語の名詞のシソーラス距離の最小値」と「前文に含まれる名詞と後文に含まれる名詞のシソーラス距離の最小値」について考えた．主語に注目したのは，前文の主題と後文の主題が変わらなければ文の展開としてさほど変化していないと考えたからである．

3.1.3 文節パタン

ここでは，文節のパタンを付与する．文節パタンとは，次のような例文があるとき

- 太郎は二郎に花をプレゼントした

次のように文節に別れる．

1. <PERSON> 太郎 </PERSON> は
2. <PERSON> 二郎 </PERSON> に
3. 花を
4. プレゼントした。

この中でも文節ごとに文節末が助詞，助動詞，すべての品詞で「非自立」であるものと動詞の「ある」についてのみ抽出し，名詞ならばその属性（PERSONやLOCATIONなど）それ以外の品詞をワイルドカードにしたものを本研究の属性とする．さきほどの例では以下のような文節パタンが出来る．

1. PERSONは
2. PERSONに
3. *を
4. *した。

3.1.4 係り受けパタン情報

文献[1]において接続関係を推定することに重要な属性である係り受けのパタン情報も用いた．これは前節で述べたものに係り受け情報を付与したものである係り受けのパタン情報とは，文を係り受け解析しその文のつながりをパタン化したものである．具体的には以下の手順で求める．

1. 文を係り受け解析し，各文節ごとに文節末が助詞，助動詞，すべての品詞で「非自立」であるものと動詞の「ある」についてのみ抽出する．その他の品詞についてはワイルドカードにする．
2. 上記で求めた文節パタンのすべての係り受けの組み合わせを求め，それらを構文パタンとする．

この係り受けパタンの長さが2から4のものを属性として用いる．

3.2 ベクトルの構築

前節で述べた属性を実際に学習するには，各単位によるベクトルを構築する必要がある．そのため，ここでは処理の流れを説明する．

1. 接続詞を挟む前後 2 文を情報として入手する
2. 文対ごとに形態素解析，係り受け解析を行う
3. 前節で述べた属性を抽出する

なお形態素解析には MeCab[5]，係り受け解析には CaboCha[6]を用いる．

[5]http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[6]http://code.google.com/p/cabocha/

4 評価実験

4.1 使用する分類器

4.1.1 決定木

ここでは，決定木学習アルゴリズムの一つである C4.5 アルゴリズム [8] について述べる．

決定木とは，予測モデルであり、ある事項に対する観察結果から、その事項の目標値に関する結論を導くものである．C4.5 では中間ノードの選定基準に以下に示す情報利得比を用いる．

情報利得比は情報利得と分割情報量によって算出される．分類するクラスが $C = \{c_1, c_2, \cdots, c_t\}$ であるとき，クラス $c_i \in C$ のサンプル数の全体に対する割合を $p(c_i)$ とすると，エントロピー $H(C)$ は以下の式となる．

$$H(C) = -\sum_{i=1}^{t} p(c_i) \log_2 p(c_i) \tag{8}$$

また，属性 $D \in \mathbf{D} = D_1, D_2, \cdots, D_N$ が名義属性 $D = \{d_1, d_2, \cdots, d_u\}$ であるとき，属性値 $d_j \in D$ でのサンプル数の全体に対する割合を $p(d_j)$，属性値 D_j・クラス c_i のサンプル数の全体に対する割合を $p(d_j, c_i)$ とすると条件付きエントロピー $H(C|D)$ は以下の式となる．

$$H(C|D) = -\sum_{j=1}^{u} \sum_{i=1}^{t} p(d_j, c_i) \log_2 \frac{p(d_j, c_i)}{p(d_j)} \tag{9}$$

情報利得 $G(C; D)$ は以下の式となる．

$$G(C; D) = H(C) - H(C|D) \tag{10}$$

また，属性 D における分割情報量 $I(D)$ は以下の式となる．

$$I(D) = -\sum_{j=1}^{u} p(d_j) \log_2 p(d_j) \tag{11}$$

これらを用いて属性 D における情報利得比 $IGR(C; D)$ は，

$$IGR(C; D) = \frac{G(C; D)}{I(D)} \tag{12}$$

となる．ここでは，全ての属性について情報利得比を求め，その中で最も大きい情報利得比を持つ属性を，決定木の中間ノードに用いる属性の第 1 候補とする．なお，属性が数値属性の場合には，各属性内に固有で情報利得比が最大になるような境界値を設定することで，名義属性と同様に扱うことが出来る．

学習をすすめていくと学習データに対して細かく条件分岐し，学習データ中の例外的な値や誤りに対して過度に適合してしまうという問題がある．これを過学習と呼ぶ．過学習した決定木は，未知データに対しては予測精度が悪化するため回避する必要がある．これを枝刈りといい，pessimistic pruning を用いている．これは終端ノードに含まれる n 個のデータを母集団から取り出したサンプルとみなし，そのサンプルのエラー率から母集団のエラー率を統計的に推定し，その値に基づいて枝刈りを行う．

4.1.2 SVM

SVM は二値分類のための教師あり学習アルゴリズムである．概念図を図 3 に示す．

学習データは（13）式のベクトルとして表すことができる．

$$(\mathbf{x_1}, y_1), \cdots, (\mathbf{x_u}, y_u), \quad \mathbf{x_j} \in \mathbf{R^n}, \quad y_j \in \{+1, -1\} \tag{13}$$

$\mathbf{x}_j$ を各学習データの特徴ベクトル，y_j を学習データ j が正例であるときに $+1$，負例であるときに -1 となる教師信号とする．

4.1.2.1 最大マージンクラス分類器

SVM は正例・負例間の最小距離（マージン）が最大となるような分離平面を決定する．マージンに対応する学習データは Support Vector と呼ばれる．

マージンは

$$\min_j \frac{|\mathbf{w} \cdot \mathbf{x}_j + b|}{||\mathbf{w}||} \tag{14}$$

となる．$\mathbf{w}, \mathbf{b}$ は，定数倍しても表現する超平面は変わらない冗長性がある．そこで，以下の制約を加える．

$$\min_j |\mathbf{w} \cdot \mathbf{x}_j + b| = 1 \tag{15}$$

このとき，次のように定式化される．

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 \tag{16}$$

$$\text{such that} \quad y_j(\mathbf{w} \cdot \mathbf{x}_j + b) \geq 1 \tag{17}$$

制約条件は，この超平面が学習事例を完全に識別することを示している．

Lagrange 未定乗数法を用いると，Lagrange 関数 $L(\mathbf{w}, b, \alpha)$ は，

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}||\mathbf{w}||^2 + \sum_{j=1}^{u} \alpha_j (1 - \xi - y_j(\mathbf{w} \cdot \mathbf{x}_j + b)) \tag{18}$$

となる．ここで，α_j は Lagrange 乗数である．したがって，次式が成り立つ．

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{j=1}^{u} \alpha_j y_j x_j = 0 \tag{19}$$

$$\frac{\partial L}{\partial b} = -\sum_{j=1}^{u} \alpha_j y_j = 0 \tag{20}$$

より，

$$\sum_{j=1}^{u} \alpha_j y_j = 0 \tag{21}$$

$$\mathbf{w} = \sum_{j=1}^{u} \alpha_j y_j x_j \tag{22}$$

の関係が得られる．よって，

$$\max_{\alpha} \sum_{j=1}^{u} \alpha_j - \frac{1}{2} \sum_{i,j=1}^{u} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x} \tag{23}$$

$$\text{such that} \quad \alpha_j \geq 0, \sum_{j=1}^{u} \alpha_j y_j = 0 \tag{24}$$

となり，多くの高速化手法の知られた α のみの凸 2 次計画問題に帰着できる．

4.1.2.2　ソフトマージン最適化

しかし，前節のモデルはノイズに対して敏感であるため，現実問題には適していない．そこで制約条件を緩める（ソフトマージン）ことを考える．このとき以下のように定式化される．

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}||\mathbf{w}||^2 + C \sum_{j=1}^{u} \xi_j \tag{25}$$

$$\text{such that} \quad y_j(\mathbf{w} \cdot \mathbf{x}_j + b) \geq 1 - \xi_j \tag{26}$$

ここで，C はどこまで制約条件を緩めるかを指定するパラメータであり，実験的に決められる．

同様に，Lagrange 関数 $L(\mathbf{w}, b, \alpha, \beta)$ は，

$$L(\mathbf{w}, b, \alpha, \beta) = \frac{1}{2}||\mathbf{w}||^2 + C \sum_{j=1}^{u} \xi_j + \sum_{j=1}^{u} \alpha_j (1 - \xi_j - y_j(\mathbf{w} \cdot \mathbf{x}_j + b)) - \sum_{j=1}^{u} \beta_j \xi_j \tag{27}$$

となる．ここで，α_j,β_j は Lagrange 乗数である．よって，

$$\max_{\mathbf{alpha}} \sum_{j=1}^{u} \alpha_j - \frac{1}{2} \sum_{i,j=1}^{u} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x} \tag{28}$$

$$\text{such that} \quad 0 \leq \alpha_j \leq C, \sum_{j=1}^{u} \alpha_j y_j = 0 \tag{29}$$

となり，同様に α のみの凸 2 次計画問題に帰着できる．

（22）式の関係から判別関数は以下のようになる．

$$f(\mathbf{x}) = \mathrm{sgn}(g(\mathbf{x})) = \sum_{j=1}^{u} \alpha_j y_j (\mathbf{x}_j \cdot \mathbf{x}) + b \tag{30}$$

また，Karush-Kuhn-Tucker 条件から，（29）式の最適解は次の条件を満たす．

$$\alpha_j = 0 \quad \Rightarrow \quad y_j g(\mathbf{x}_j) \geq 1 \tag{31}$$

$$0 \leq \alpha_j \leq C \quad \Rightarrow \quad y_j g(\mathbf{x}_j) = 1 \tag{32}$$

$$\alpha_j = C \quad \Rightarrow \quad y_j g(\mathbf{x}_j) \leq 1 \tag{33}$$

$\alpha_j \neq 0$ となる $\mathbf{x}_j$ は Support Vector と呼ばれ，判別関数は Support Vector によってのみ記述される．ソフトマージンの場合，$\alpha_j > 0$ であるような学習データはすべて Supoort Vector と呼ばれる．

また，SVM は学習問題，判別関数と学習データの内積で記述されるため，（35）式のように内積を Kernel 関数で置き換えることで非線形の分離平面を実現できる特徴がある．

$$g(\mathbf{x}) = \sum_{j=1}^{u} \alpha_j y_j K(\mathbf{x}_j \cdot \mathbf{x}) + b \tag{34}$$

$$K(\mathbf{x} \cdot \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^d \tag{35}$$

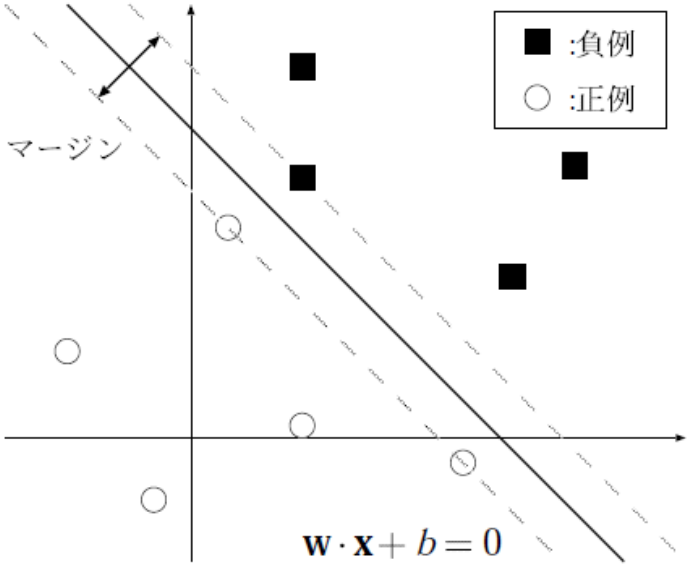


図 3: SVM の概念図

4.2　学習データの収集とその調査

必須箇所と推奨箇所について以下に述べる方法で調査を行った．調査に使用した文章は，毎日新聞コーパスから接続詞のある文の前後をランダムに選択した文章である．また全ての調査において同じ記事を用いた．

4.2.1　必須箇所

ここでは，接続詞がなければ意味が通じなくなる箇所を収集する．まず，接続詞を隠した 2 文を被験者に提示して，隠された接続詞を推測させた．その後，実際に用いられている接続詞を提示して，推測してもらった接続詞と一致したかどうかを判断してもらった．また客観的にデータを収集するため被験者を 3 人割り当て，各被験者による判定の相違を多数決により決定した．

被験者は成人 21 名であり，接続詞を挟む前後文を 1 項目として一人あたり 100 項目の調査を行った．調査の結果，収集されたデータは,「判断と同じ＝接続詞が必須ではない」が 521 件,「判断と異なる＝接続詞が必須」が 85 件の計 606 件であった．

● 調査

ここでは，推奨箇所における収集データについてどのような傾向があったかについて調査した．

■ 接続詞ごとの集計結果

各接続詞ごとにどのように分類されたかを表 11 に示す．また表中の対応はない方が良い:ok，あった方が良い:half，判断と違った:ng としている．

この結果を見ると，判断と違ったと答えた人は予想以上に少ないことがわかった．これは人は前後文との接続関係をたった 2 文でも判断することができることを示していると考えられる．

一方，ない方が良いと答えたものはほとんどないという結果となった．これは，文と文の繋がりで接続詞が必要かどうかは 2 文という少ない情報量では接続詞が必要である場合が多いという事が分かる．

つまり，接続詞が省略出来るかどうかは 2 文では決まらず，複数文で決まる可能性がある．

表 11: 合計数 10 以上の接続詞についての結果

接続詞の種類	ok	half	ng	合計
しかし	0	116	7	123
また	6	65	14	85
一方	3	73	8	84
だが	4	72	5	81
ただ	1	40	15	56
でも	2	29	4	35
そして	2	18	4	24
ところが	0	16	0	16
そこで	0	11	0	11
それでも	0	10	0	10

■ 接続関係における偏り

次に接続関係ごとに結果を表 12 に示す.

表 12: 接続関係ごとの結果（必須箇所）

接続詞の種類	ok	half	ng	合計
逆説	7(2.5%)	251(90.9%)	18(6.5%)	276
累加	11(5.7%)	151(73.7%)	43(20.9%)	205
並列	4(4.4%)	75(83.3%)	11(12.2%)	90
因果	1(2.9%)	30(85.7%)	4(11.4%)	35
例示	2(14.3%)	11(78.6%)	1(7.1%)	14
合計	25(4.0%)	518(83.5%)	77(12.4%)	620

この結果を見ると，判断と違ったで多かったものは累加であった（接続詞の対応表は 2 節の表 3 を参照）．これは，累加の「また」「そして」と並列の「または」「もしくは」や累加の「ただ」と逆説の「しかし」と間違えることなどが考えられる.

4.2.2　推奨箇所

4.2.2.1　文脈なし

ここでは，文としての繋がりが自然であるかを考慮した上で接続詞が必要である箇所を収集する．接続詞を挟む前後 2 文を表示して接続詞が「省略出来るか」の判定をしてもらった．接続詞が省略できるかどうかは主観に左右されることが多いため，各前後文に対して 2 人の被験者を割り当てて，判断の一致を考慮した．

被験者は成人 22 名であり，接続詞を挟む前後文を 1 項目として一人あたり 100 項目の調査を行った．2 名の判定が一致した項目を有効なデータとしてコーパスに追加した．その結果，収集されたデータは「省略できる箇所」が 231 件,「省略できない箇所」が 590 件の計 821 件であった．

● 調査

ここでは，推奨箇所 : 文脈なしにおける収集データについてどのような傾向があったかについて調査した．

■ 接続詞ごとの集計結果

表 13 に 10 個以上の接続詞について結果を示す．ここで，Positive は省略可能，Half は人によって結果が異なったもの，Negative は省略出来ないと判断したものである．この表をみると，種類に関係なく省略が可能であることがわかる．やはり省略出来ないと判断するほうに偏りはあるが，出来ないものや人によって意見がわかれるものも多数存在する．

この結果から分かることは，接続詞単体だけでは省略可能かどうかは判断出来ないという事である．つまり，本研究の意義は示されていると考えられる．

■ 接続関係における偏り

次に接続関係ごとに集計した結果を表 14 に示す．接続関係については，2 節の表 3 のように分類している．

この表を見ても，特に接続関係によって偏りがあるとは言えず，平均的に省略が出来るものと出来ないものがあることが分かる．

■ 推奨箇所と必須箇所を比較した考察

推奨箇所では最初に提示して省略出来るかどうかを判断してもらった．必須箇所では逆に最初は提示せず省略出来るかどうかを判断してもらった．

表 13: 合計数 10 以上の接続詞についての結果

接続詞の種類	Positive	Half	Negative	合計
しかし	38	51	106	195
また	37	40	84	161
一方	29	43	81	153
だが	29	32	82	143
ただ	16	32	57	105
でも	15	10	27	52
そして	13	9	22	44
ところが	5	8	13	26
例えば	2	7	10	19
それでも	2	5	10	17
まずは	2	6	6	14
そこで	2	7	4	13
しかも	3	5	4	12
だから	4	2	5	11
ただし	2	1	7	10
そのうえで	3	5	2	10

その結果から，本研究における接続詞が省略が出来る状況を定義すると，必須箇所で省略が可能と判断され，かつ必須箇所でない方が良いとなったものに対してのみ省略することができるとする．

推奨箇所の関係（省略可能:yes，省略不可能:None）と必須箇所の関係 (ない方が良い:ok，あった方が良い:half，判断と違った:ng, 判別不能:muri) をまとめたものを表 15 に示す．また，結果の集計における上記の関係の決定は推奨箇所では 2 人の被験者の判断が一緒であったものでかつ，必須箇所では 3 人中 2 人以上の被験者の判断が一緒であったものが被ったものに対してのみで行った．その合計は 284 文である．

この結果をみてもわかるが，推奨箇所で省略可能と判断されかつ必須箇所で判断と一緒であったもの（つまり yes-ok と yes-half）の合計は 64 文しかなく，共に被っていた 700 文に対して，非常に少ない結果となっている．

しかし，推奨箇所で省略出来ないと判断したものの，実際には判断が正しかったが

表 14: 接続関係ごとの結果（推奨箇所：文脈なし）

接続関係の種類	Positive	Half	Negative	合計
逆説	94	110	249	453
累加	82	106	207	395
並列	30	45	88	163
因果	16	11	21	48
例示	6	7	12	25
転換	3	0	13	16

表 15: 推奨箇所と必須箇所の判断結果のまとめ (合計 284 文)

必須箇所 / 推奨箇所	省略可能 (yes)	省略不可能 (None)
ない方が良い (ok)	3(1.1%)	11(3.8%)
あった方が良い (half)	61(21.4%)	159(56.0%)
判断と違った (ng)	7(2.4%)	28(9.9%)
判別不能 (muri)	4(1.4%)	11(3.8%)

あった方が良いとなっているものについては (つまり None-half)，159 文あるため人はそれほど意味を捉え違えることはないとも考えられる．それは判断と違った (ng) の合計が 35 文であることを考慮すれば約 12%であることからも言えると考えられる．

4.2.3 調査データの信用性

上記で述べた調査は大学生を中心に行ったため，被験者の文章作成能力を測った．文章作成能力とは，文法的に正しい日本語を使って文章を作成する能力である．ここでは被験者に対して以下の 6 つの項目を用いてアンケートを実施し，文章能力を測定した．

1. 1 週間における平均読書（新聞を含む）冊数（ページ数単位）

2. 1 週間における平均読書（新聞を含む）時間（時間単位）

3. 図書館の貸出数（冊数）（内，専門書の数も聞く）

4. 一週間の平均文章作成時間（論文，SNS）

5. 一週間の平均文章作成文字数（論文，SNS）

これらのアンケートから，文章作成能力が高い人と低い人に分類した．分類方法は以下のようになっている．

1. 各人各項目のデータを d_{ij} とする

2. 各項目 j ごとの中央値 M_j を求める

3. 中央値 M_j が各データ d_{ij} よりも高いか低いかで $b_i = \{0, 1\}$ を求める

4. 各人 i の b_i の合計を Sum_i とする

5. Sum_i の中央値を $Median$ とする

6. 中央値 $Median$ が各人の合計 Sum_i よりも高いか低いかで最終的な作成能力 $Brain_i = \{0, 1\}$ を求める

この分類において，文章作成能力が高い人に分類された被験者だけのデータによって構成されたコーパスを用意した．

5　結果と考察

学習を行う上で，調査で得られた学習データの正例と負例の個数が異なるため，正例よりも数が多い負例をランダムに選択し同数にして学習を行った．必須箇所，推奨箇所の分類器ともに 10 分割交差検定により評価を行った．

分類性能の評価としては再現率，適合率，F 値を用いた．ここで，再現率は完全性を評価するための尺度であり，分類器によって正しく判断された正例データ数を正例データ数で割ったものである．適合率は正確性を評価するための尺度であり，分類器によって正しく判断された正例データ数を分類器によって正例と判断されたデータ数で割ったものである．F 値は再現率と適合率の調和平均である．

分割された各テストデータの各評価値の平均を全体の評価値とし，負例データを選択しなおして行った 10 回の実験の平均を最終的な評価値とした．その結果を表 16，17 に示す．

5.1　必須箇所

表 16: 必須箇所の分類器における分類性能

使用した学習手法	評価基準		
	適合率	再現率	F 値
決定木	0.78	0.45	0.56
SVM	0.57	0.63	0.60

表 16 を見ると，決定木では適合率が高い結果から，分類器が接続詞が必要である箇所と判断するものに対しては正しい結果を得られると言えるが，再現率が低い結果から，元々接続詞が必要である箇所を正しく判断することが出来なかった．一方，SVM では決定木よりは再現率が高いものの，適合率は低くなった．本研究の目的は文章添削システムの構築であるため，分類器が接続詞が必要と判定した精度，適合率が高い方が望ましい．また，ランダムに選択するベースライン（0.5）よりも高い結果を得られている．そのため，添削システムとして用いるならば SVM よりも決定木を用いたほうが望ましいことがわかった．

5.2 推奨箇所

表 17: 推奨箇所の分類器における分類性能

使用した学習手法	評価基準		
	適合率	再現率	F 値
決定木	0.81	0.49	0.57
SVM+新聞著者を信用したデータ	0.89	0.94	0.92

表 17 を見ると，決定木では必須箇所同様に「省略できない」と分類されたものに関しては正しく分類出来ているものの，本来「省略できない」データを適切に分類出来ていない結果となった．これは，適合率がベースライン（50%）よりも高い結果から有効である事がわかった．

一方，アンケート調査に頼らず新聞記事において接続詞がある文と無い文を用いた収集も行った．正例を接続詞がある箇所，負例を接続詞がない箇所としそれぞれ 900 件収集した．これは，新聞記事を作成している著者はプロフェッショナルであるため，適切な接続詞を入れていると考えられるからである．新聞著者を信用した実験では，とても良い結果となった．これは，新聞著者のような文章作成能力が高い人は接続詞が必要となる箇所を決める方法が明確に存在すると考えられる．

5.3 信用性を考慮した実験

表 18: 信用性を考慮した必須箇所，推奨箇所の分類器における分類性能

分類箇所+使用した学習手法	評価基準		
	適合率	再現率	F 値
必須箇所+決定木	0.84	0.47	0.60
必須箇所+SVM	0.56	0.57	0.56
推奨箇所+決定木	0.83	0.46	0.58

ここでは，被験者の文章作成能力が高ければ 5.1，5.2 節での評価実験よりも適合率，再現率ともに高くなる可能性がある．必須箇所では，決定木と SVM で異なる結果となった．決定木では信用性を考慮すると適合率，再現率ともに上昇した．これは文章作成能力がある人では接続詞が必要となる箇所に対して共通のルールのようなものがある可能性を含んでいる．SVM では信用性を考慮すると適合率，再現率ともに悪くなった．これは，予想に反した結果であるがその原因として SVM ではある程度の学習量が必要であるが，信用性を考慮したことによりデータ数が少なくなったことから正しく学習できなかったと考えられる．

一方，推奨箇所では，信用性を考慮しても大きな精度の変化が見られなかった．この結果は例えば推奨箇所の調査内容は人が判断する際，曖昧な部分が数多くあるため作成能力の有無によって適合率や再現率が高くなる起因として有効に働かなかったという可能性が考えられる．

5.4 必須箇所と推奨箇所の相互関係について

表 19: 必須箇所と理想箇所の精度

理想＼必須	必ず必要	必ずしも必要ではない
出来れば必要	38.2% (13/34)	33.3% (8/24)
出来れば必要ではない	26.9% (7/26)	53.1% (17/32)

必須箇所の決定木での分類器と推奨箇所の決定木での分類器を用いて各データセットにおける判定の違いの分析を行った.

分析では，評価用データとして 4.2.1 節の必須箇所における調査結果と 4.2.2 節の推奨箇所における調査結果が同じ記事においてあるものだけを用いて行う．判定するデータは合計 116 件である．ただし，そのなかでも調査データ時点での誤り，接続詞が必ずしも必要としないが出来れば必要ではない（必須箇所にて正が正解かつ理想ラインでは正が正解）が 24 件あった．結果を表 19 に示す．ここで，先に述べた調査データの誤りを考慮した正解数は 26.9%(7/26) であった.

表 19 をみると，ランダムに判定を行えば 25 %であるのでそれよりは多くなった．正と正，負と負に関しては精度が高いため正しく分類されていることがわかった．このことから，必須箇所と理想箇所において相互に違う特徴を持った分類器が正しく分類できていることがわかる.

ここで特に，判定が異なる部分である必須箇所では負（必要ない）であったものの，推奨箇所では必要となった結果は以下のような文があった.

- 継続は力なり、という心境だ」と話した。また、「医師としての経験を生かし、自分の体が宇宙滞在によってどう変わるのかを確かめ、誰もが宇宙に行ける時代の準備に貢献したい」と抱負を述べた。

- ホンダの参戦費用は年間約５００億円といわれ、ほとんどを自社で賄っていた。しかし販売の落ち込みと円高で、０８年度下期は１９００億円の赤字の見込み。

このような文ではたしかに意味を考えれば，この 2 つは接続詞が必ずしも必要ないが理想的には必要であることがわかる．

一方，

- また、宗匠（師匠）の意向通りに作句していた当時の流行に抵抗して、新たな俳句を作り上げようという子規の新たな試みが表れて、漱石の句にも同様の斬新さがうかがえる。ただ、「老成していた」との感想も誤りではないと渡部さんは話した。

この文では，何回か読みなおせば後文における「〜との〜も〜ではない。」の部分から「ただ」になることがわかるが，すぐに理解するのは厳しい．

このように実際に判定されたデータをみると，判定が非常に曖昧なものが多く見受けられた．

5.5　必須箇所を用いた要旨添削事例

ここでは，本研究の要旨における第 1 節である「はじめに」を必須箇所の分類器を用いて，実際に添削した結果について述べる．要旨における「はじめに」は全 13 文あるため，判定する文間は 12 箇所である．実際の「はじめに」は付録 D に載せる．

実際に接続詞が用いられていて，かつ分類器が必須と判断した箇所を接続詞を赤文字とし，実際に接続詞が用いられているが，必須であると判断されなかった箇所の接続詞を青文字とした．この箇所において，接続詞が必要であるかを 5.3 節での信用性を考慮した必須箇所の分類器を用いて評価実験を行った．その結果を表 20 に示す．

表 20: 本研究の要旨における第 1 節の詳細結果

		実際の正解		合計
		接続詞がある	接続詞がない	
システムの出力	接続詞が必要	3	5	8
	接続詞が必要ない	2	2	4
合計		5	7	12

この結果から，適合率 37.5%，再現率 60%, F 値 46.1%となった．ここでその詳細な結果をみると，実際の正解において「接続がない」文間にはその 2 文の間に「その際」,「そのため」などの文献 [9] らが接続詞的機能をもつ語句があった．それらを正解（接続語句が必要）とし判定すると表 21 のようになる．

表 21: 本研究の要旨における第 1 節の詳細結果

		実際の正解		合計
		接続詞がある	接続詞がない	
システムの出力	接続詞が必要	5	3	8
	接続詞が必要ない	2	2	4
合計		7	5	12

この結果から，適合率 62.5%，再現率 71.4%, F 値 66.6%となった．そのため，必

須箇所の分類器によって接続詞の必要箇所を導けていることがわかり，本研究の有用性を確認できる．

具体的に文を見ると，2段落目4行目では，接続詞「しかし」がなくても比較的容易に接続関係が「逆接」とわかるため接続詞は必ずしも必要ないことがわかる．この結果から接続詞が必要な箇所を正しく導けていると言える．赤文字の最終段落の「よって」の接続詞は必須ではないと考えられる．後文にある「これらの」が前文までのことを指しているため比較的容易に「因果」であると導けるからである．そのため，改段落などの文構造における技術的要素や意味情報を考慮する必要がある．

5.6 有効な属性の分析

ここでは，3.1 節で述べた本研究で提案した属性のどの要素が重要であったかについて述べる．分析方法として各属性を抜くことでどのような変化が見られるかと一つの木に注目する方法の 2 種類を考える．

5.6.1 属性削減による分析

表 22 に使用する属性の種類を，それに対応した決定木における分類器の結果を表 23 に示す．なお必須箇所，推奨箇所ともに使用した学習データは信用性を考慮したデータである．

表 22: 組み合わせる属性一覧（使用する：○，使用しない：空欄）

種類	助詞，助動詞	繰り返し語句	シソーラス距離	文節パタン，係り受けパタン
属性集合 1	○	○	○	○
属性集合 2	○	○	○	
属性集合 3	○	○		○
属性集合 4	○	○		

表 23: 各属性を使用した必須，推奨箇所の分類性能

箇所	組み合わせる属性	評価尺度		
		適合率	再現率	F 値
必須	属性集合 1	0.84	0.47	0.60
必須	属性集合 2	0.60	0.53	0.54
必須	属性集合 3	0.80	0.45	0.57
必須	属性集合 4	0.61	0.54	0.57
推奨	属性集合 1	0.83	0.46	0.58
推奨	属性集合 2	0.68	0.50	0.56
推奨	属性集合 3	0.83	0.45	0.57
推奨	属性集合 4	0.59	0.52	0.55

結果では最も属性を考慮しない属性集合4を基準に考える．属性集合3をみると，必須箇所では文節パタン，係り受けパタンを追加すると属性集合4に比べて適合率が高くなった．また，属性集合1をみるとシソーラス距離では文節パタン，係り受けパタンと組み合わせると適合率が高くなるものの，属性集合2のように文節パタン，係り受けパタンがなければ両方削減した結果とほとんど変わらない結果であった．

そのため，文節パタンと係り受けパタンは接続詞の必要箇所の分類性能を上げる要因であったと考えられる．そして，シソーラス距離の情報は文節パタンと係り受けパタンとを組み合わせることで分類性能を上げる要因であったと考えられる．

次に，推奨箇所でも必須箇所の場合と同様に最も属性を考慮しない属性集合4を基準に考える．属性集合3をみると文節パタン，係り受けパタンを追加すると属性集合4に比べて適合率が上昇しているのがわかる．また，属性集合2をみるとシソーラス距離を追加すると属性集合4に比べて適合率が上昇しているのがわかる．しかし，これらを組み合わせた属性集合4では属性集合3と比較しても精度は変わらなかった．そのため，文節パタン，係り受けパタン，シソーラス距離は推奨箇所の分類性能を上げる要因であるものの，それらを組み合わせるとシソーラス距離の影響はほとんどなくなってしまったことが分かる．これは，文節パタンと係り受けパタンの影響がとても大きいことが原因であると考えられる．

5.6.2 決定木における重要な属性

本研究の最終的な目標は文章添削システムであるため，必須箇所の分類器が重要となる．学習された決定木の葉の部分に着目することでそれぞれの属性によって「接続詞が必要」と「接続詞が必要ない」がどの程度分類されたかを分析し，分類に重要な属性について考察する．この方法により，今回の分類問題である接続詞が「接続詞が必要」「接続詞が必ずしもいらない」それぞれに有効である属性が何なのかがわかる．ここでは，全ての評価実験の中で最も精度が高かった必須箇所における信用性を考慮した分類器を例として示す．

分析の結果を図4に示す．後文にパタン情報の「*も」[7]や前文にパタン情報「わ

[7]ワイルドカードがある場合：前の単語が助詞，助動詞，すべての品詞で「非自立」，動詞「ある」以外のもの

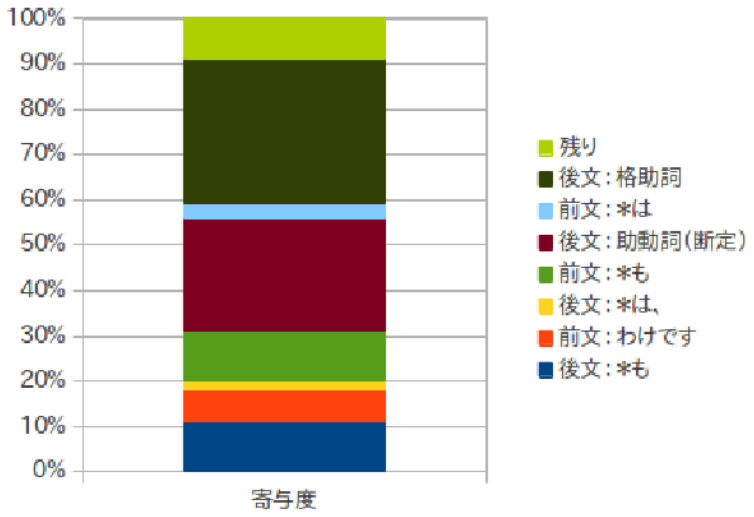


図 4: 最も精度が高かった決定木の詳細結果

けです」が存在するときは「接続詞が必要」と判断されやすい結果になった．これらのパタンは文間にある様々な接続関係によって話題を変えるときに利用されるからと考えられる．また，前文に「*も、」や後文に「助動詞：断定」が存在するときは「接続詞が必要ではない」と判断されやすい結果であった．これらが使われるのは文の話題が切り替わる接続詞が比較的容易に想像できることから重要な属性であると考えられる．

ワイルドカードがない場合：前の単語がすべての品詞

5.7 文章添削システムに向けて

表 24: 本研究を適用した若山らの手法 [1] の分類性能

接続関係	本研究適用前			本研究適用後		
	適合率	再現率	F 値	適合率	再現率	F 値
並列	0.57	0.17	0.26	0.60	0.30	0.40
例示	0.00	0.00	0.00	0.00	0.00	0.00
因果	0.13	0.33	0.19	0.13	0.33	0.18
累加	0.60	0.17	0.27	0.71	0.20	0.31
転換	0.05	0.67	0.10	0.06	0.50	0.11
逆接	0.44	0.46	0.45	0.48	0.45	0.46
全体	0.30	0.30	0.30	0.33	0.30	0.31

ここでは，我々の最終目標である文章添削システムに向けてその一つの推定要素である「論理関係の適切さ」の判別について述べる．添削システムとしての最終的な課題は文間に入る接続詞（もしくは接続関係）が正しく推定できることであるため，若山らの手法 [1] が本研究を適用することによってどの程度向上したかを考える．実験方法はまず，入力された文章における全ての文間に対し必須箇所における信用性を考慮したデータを用いた決定木の分類器により接続詞が必要となる箇所を判定する．次に，接続詞が必要と判定された文間に対して，若山らの手法により接続詞を推定する．

評価方法として 4.2.1 節にて収集したデータを用いて判断を行った．本研究を適用する前後での若山らの手法の分類性能の結果を表 24 に示す．この結果から「並列」,「累加」,「転換」,「逆接」に関しては分類性能が上昇したため，本研究が有効に働いたと考えられる．

具体的にみると「累加」の適合率が約 10%近く上昇しているのがわかる．これは，本研究を適用することで推奨箇所における接続詞を削減したことから，接続関係を同定するのに曖昧な部分が軽減したことが示唆される．また,「並列」に関しては再現率が上昇しているのがわかる．これも本研究を適用したことにより接続詞がなくても「並列」の関係を導けるものの，若山らの手法ではとれなかったものが上手く

軽減出来ていることが分かる．

一方,「例示」「因果」については精度がほとんど変わらなかった．そのためこれらの接続関係に特化した接続詞の必要箇所の推定や「例示」や「因果」の接続関係の同定の分類性能をより向上する必要がある．

6　おわりに

本研究では文章の自動添削という観点から，接続詞が必要とされる箇所を推定するために品詞や文節パタン，シソーラス距離などの属性を用いて機械学習手法により分類する手法を提案した．

評価実験の結果，適合率がランダムに選択するよりも高く，本手法が接続詞の必要箇所の判定に有効であるとわかった．

今後の課題として，以下のようなものが挙げられる．

- 意味的な情報の属性への追加
- 各接続関係に特化した手法の開発
- 学習に使用したデータの大規模化

意味的な情報は本研究ではシソーラス距離が該当するが，より詳細な分析を行うことで精度の向上が期待できる．各接続関係に特化した手法は，6 種類ある接続関係ごとに独立した接続詞の必要箇所の分類器を構築することでより全体の精度が期待できる．ただし，接続関係の同定の研究をより精度を向上させる必要がある．

参考文献

[1] 若山裕介，内海彰：SVM を用いた接続関係の同定，人工知能学会第 26 回全国大会論文集（2012）.

[2] 柴崎秀子，原信一郎：12 学年を難易尺度とする日本語リーダビリティ判定式，計量国語学，vol.27，No.6，pp.215-232 (2010).

[3] 佐藤理史，均衡コーパスを規範とするテキスト難易度測定，情報処理学会論文誌，vol.52，No.4，pp.1777-1789 (2011).

[4] 建石由佳，小野芳彦，山田尚勇：日本文の読みやすさの評価式，情報処理学会文書処理とヒューマンインタフェース，Vol.18，No.1，pp.1–8 (1988).

[5] 石岡恒憲：小論文およびエッセイの自動評価採点における研究動向, 人工知能学会誌，vol.23, No.1，pp.17–24 (2009).

[6] 山本和英，斎藤真美：用例利用型による文間接続関係の同定, 自然言語処理，Vol.15，No.3, pp.21–51 (2008).

[7] Emily Pitler, Annie Louis and Ani Nenkova : Automatic sense prediction for implicit discourse relations in text, *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pp.683–691 (2009).

[8] J. R. Quinlan : C4.5: Programs for Machine Learning, *Morgan Kaufmann* (1993).

[9] 市川 孝：国語教育のための文章論概説，教育出版 (1978).

[10] 此島正年 : 国語助動詞の研究 体系と歴史，桜楓社 (1973).

[11] 松吉俊，佐藤理史：文体と難易度を制御可能な日本語機能表現の言い換え，自然言語処理，Vol.15，No.2，pp.75–99 (2008).

謝辞

本研究を行うにあたり，丁寧にご指導くださった内海彰先生に深く御礼申し上げます．研究室の同期である小野くんには考察や研究活動にて大変お世話になったので深く御礼を申し上げます．研究室の聴講生である近藤さんには実装などで大変参考になったので深く御礼を申し上げます．

また，学習データの収集やアンケートに答えて下さった方々に深く御礼を申し上げます．

A　学習データの収集に用いた実験の詳細

A.1　必須箇所

収集する際における説明文は以下のようになっている.

説明
これから、文の接続詞が省略できるかどうかの判定実験を行ないます。
出題形式は，もともと接続詞がある文を伏せた状態で提示します。
次に，その伏せた接続詞がどのような関係なのかを想像してください。
その後，伏せていた接続詞を表示します。

想像した接続詞と関係が一緒であり、接続詞がない方が自然な場合は「ない方が良い」にチェックしてください。
想像と一緒であったが接続詞がある方が自然である場合は「あった方が良い」にチェックしてください。
判断と異なっていた場合は「判断と違った」にチェックしてください。
また、文章が明らかにおかしい場合や接続詞がどう見ても変など、処理のミスを疑える場合は「判別不能」にチェックしてください。

実験における注意事項：
1．この実験は専門的な知識は要求されていないので、気楽に受けてください
2．実験は全部で 100 文あります。
3．各文に対し、まず「前文」と「後文」を見て隠れている「接続詞」を想像してください。
4．次に、「見る」ボタンをクリックし想像した接続詞との関係があっていたかを判断してください。
5．各ページ 10 件あるので，ページが終わり「次へ」のボタンを押し画面が切り替わると自動的に保存され、その時点での中断が可能です。
6．再開する時はこのページに戻り、同じ名前を入力して始めてください。

7．決してブラウザの戻るを押さないでください。

それでは、名前を入力して実験を始めてください。
よろしくお願い致します。

収集する Web ページを図 5 に示す.

No.	前文	接続詞	後文	省略可能
1	【小林祥晃】JR西日本は、地震で一定以上の揺れを観測すると、周辺の列車を緊急停止させ、線路に異常がなければ運転を再開させている。	しかし	大規模地震では点検に時間がかかり、多くの列車に影響が出るとみられる 。	正解は「しかし」 ○ ない方が良い ○ あった方がいい ○ 判断と違った ○ 判別不能
2	田辺さん方は、夫妻と4人の子供、宗一さん、政行さんの8人家族。	[???]	、奈良市大柳生町では午前2時半ごろ、造園業、谷本栄さん（60）方から出火 。	見る
3	奈良県警奈良署は谷本さんの妻節子さん（59）の父万治郎さん（101）、母タケエさん（90）とみて身元の確認を急いでいる。	[???]	午前1時35分ごろ、大阪府高石市取石4、不動産業、嶋田亮さん（46）方から出火、木造2階建て住宅約150平方メートルのうち約85平方メートルを焼き、府警高石署によると、嶋田さんの母照さん（76）が1階浴室で焼死した 。	見る
4	ガザでの戦闘長期化に伴い、国際機関などが人道状況の悪化に懸念を表明していた。	[???]	、市民の死傷は3日の地上侵攻後、増大の一途をたどり、AFP通信によると、先月27日の空爆開始以降のパレスチナ人死者は660人に上り、うち215人が子供、98人が女性と	見る

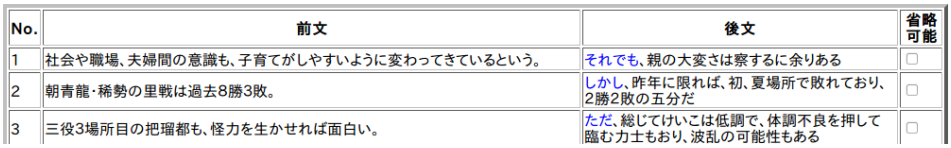
図 5: 必須箇所における実験 WEB ページ

A.2　推奨箇所

　収集する際における説明文は以下のようになっている。

説明
これから、文の接続詞が省略できるかどうかの判定実験を行ないます。
出題形式は 2 文の間に接続詞がある状態では省略できるかどうかを、接続詞がない状態ではそこに接続詞が必要かどうかを判断してもらいます。

収集する Web ページを図 6 に示す.



No.	前文	後文	省略可能
1	社会や職場、夫婦間の意識も、子育てがしやすいように変わってきているという。	それでも、親の大変さは察するに余りある	☐
2	朝青龍・稀勢の里戦は過去8勝3敗。	しかし、昨年に限れば、初、夏場所で敗れており、2勝2敗の五分だ	☐
3	三役3場所目の把瑠都も、怪力を生かせれば面白い。	ただ、総じてけいこは低調で、体調不良を押して臨む力士もおり、波乱の可能性もある	☐

図 6: 推奨箇所：文脈なしにおける実験 WEB ページ

A.3　調査データの信用性

　調査データの信用性を考慮するためのアンケート調査における説明文は以下のようになっている。

説明
これから、文章作成能力に対してのアンケートを答えてもらいます。
以下のアンケートに全て答えて送信ボタンを押してください。また、中断も可能です。その時も送信ボタンを押してください。次回入力された状態で保存されています。

注意事項:
読書とは、文庫本などの一般的な本はもちろんですが、論文や参考書などもカウントしてください。
文章の作成では、普段のレポートやゼミの資料、論文投稿を論文等として、 Twitter や Facebook、ブログの投稿などは（SNS 等）としてカウントしてください。
図書館の貸出数は大変恐縮ですが、図書館にて貸し出し履歴を参照してもらってください。本人であれば参照（印刷）が可能です。ご協力お願い致します。

収集する Web ページを図 7 に示す.

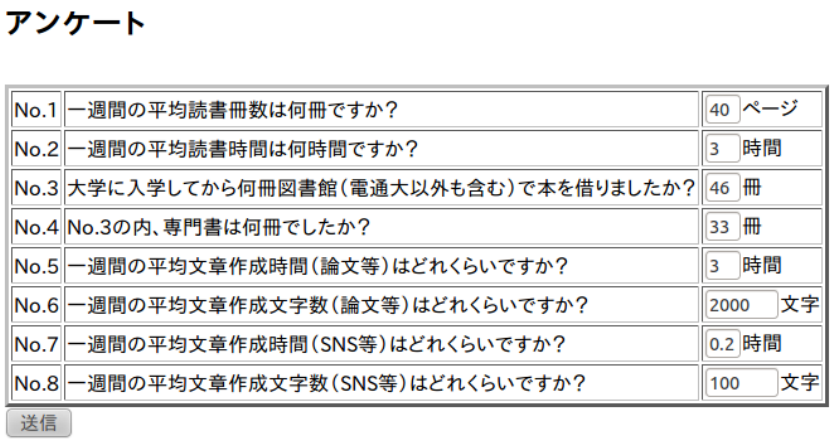


図 7: 調査データの信用性を考慮するためのアンケート調査における実験 WEB ページ

B　学習データの収集における調査例

B.1　実験 1 の判定結果の例文

省略可能 (Positive)

- 皇太子妃雅子さまは適応障害のため長期療養中ながら、主治医と相談しながら公務を少しずつ増やしており、昨年１２月９日の４５歳の誕生日には「お医者様のご指導の下で一つ一つ努力を重ねていきたい」との感想を公表した。しかし、完全な回復にはまだ時間がかかりそうだ。

- 「最高の１年」と振り返る０８年は、優勝４回で年間最多勝（７９勝）を記録し、「横綱の責任を果たせた」と喜ぶ。, しかし、さらに今年の目標を聞くと、「６場所９０日間すべて出ること」と控えめだ。

- また, 昼はさらに「口上」で市川團十郎家独特の「にらみ」を行い、「お祭り」も踊る。また夜の「白浪五人男」の弁天小僧は９８年以来。

- また, 都内多摩地区では昨年３月と６月、ホームレスの男性が襲撃される殺傷事件が他に４件発生。また、今月２日に世田谷区喜多見で殺害されたホームレスの男性は住所不定、無職、近藤繁さん（７１）と判明した。

人によって違った結果 (Half)

- 彼は１９歳にして盲学校の教師に任命され、ブライユ点字で生徒たちに教えた。しかし、盲学校は彼の点字を公式には認めなかった。

- 自民党は新年度の予算重要政策に農山漁村や商店街の活性化を盛り込み、民主党も０７年の参院選で「コミュニティーの再生」を掲げ、地域重視を看板にしている。しかし、地域崩壊の進行は中央の想定よりずっと速い。

- マシンは「走る実験室」と呼ばれ、開発された技術の多くが市販車に反映された。また、後にホンダの中枢を担う多くの技術者の育成にも貢献した。

- ６０年発表の「風流夢譚」では右翼の襲撃事件を引き起こすなど物議を醸した。また、今川焼き店を開くなど浮世離れした生き方が多くのファンをつかんだ。

省略不可能 (Negative)

- ＦＩＡのマックス・モズレー会長は、数カ月前から「このままではＦ１は持続不可能になる」と大胆なコスト削減の必要性を訴え、全チームが同じ仕様の「統一エンジン」を使う案を持ち出した。しかし参戦メーカー側は「自社技術の開発ができなければ参戦の意味がない」と反発。

- ホンダの参戦費用は年間約５００億円といわれ、ほとんどを自社で賄っていた。しかし販売の落ち込みと円高で、０８年度下期は１９００億円の赤字の見込み。

- 国内には屋久島（鹿児島県)、白神山地（青森、秋田県)、知床（北海道）の三つの自然遺産があります。また世界遺産のうち、戦争、災害、開発などで価値を損なう危険がある世界遺産は「危機遺産リスト」に登録されます。

- パレスチナ情勢に詳しいイラン人ジャーナリスト、ザイディアバディ氏は「イラン指導部にはアラブ諸国がイスラエルと結託してハマスをつぶそうとしていると映る。また、ハマスつぶしはイラン攻撃に等しいと認識している」と解説する。

B.2　必須箇所の調査と推奨箇所の調査での人が判断した結果の例文

省略可能かつない方が良い (yes-ok)

- 言いたいことを言って再び友好的な笑顔に戻る韓国人学生に怒りが募り、仲たがいはしないものの思い出すと嫌な気分になるという。実はかなり頻繁にあるケースだ。

- 【古本陽荘、近藤大介】首相は「１億円も収入がある人はもらわないのが矜持（きょうじ）」などと発言していた。だが所得制限を巡って政府・与党は迷走し、結局は「地方自治体の判断」に丸投げした。

省略可能かつあった方が良い (yes-half)

- 裁判員裁判は丸１日を費やすことが多いとみられ、日当は満額が支払われそうです。一方、裁判員の選任手続きは午前中で終わるのが通常で、裁判員に選ばれなかった候補者の日当は上限の半額程度になる見通しです。
- 広大な敷地内には、アブダビＧＰが行われるサーキットも含まれる。しかしオイルマネーに支えられた中東の好況も、原油価格の急激な下落で暗雲が漂い始めた。

省略不可能かつない方が良い (None-ok)

- そして世界不況に直面した今日である。だが積み重なった過去を一つ一つ思い起こせば、苦境を乗り切り、平成を願う心の支えも案外近くに見つかろう。
- でも、脳の老化防止には役立ちそうだと気長に続けることにしました。そして同時に「求めない」「期待しない」を、私の新たなスローガンにしました。

省略不可能かつあった方が良い (None-half)

- 水泳もドライヤーもＯＫだが、増やせる本数に限りがあり、「負担が大きく、２０００本が限界だと思う」と話す。一方、ウイッグは、不自然さをなくせば、確実に増やせて髪形も自由だ。
- 関脇・把瑠都とは取らなかったが、豪栄道や武州山を相手に、ひじを痛めている左からのど輪も繰り出したほか、新入幕の山本山の相手も。ただ立ち合い正常化を打ち出す武蔵川理事長（元横綱・三重ノ海）からは、左の手つきが不十分と注意され、神妙にうなずく場面もあった。

省略不可能かつ判断と違った (None-ng)

- 何が始まるか察した子も多いようだ。まずは男女各１０人が並べた椅子の周りに立つ。
- 公明党幹部によると、４日夜の首相と与党幹部との会合で、自民党の大島理森国対委員長が首相に対し「（党内の造反の）動きのある人に電話してください」と求め、首相が「分かった」と応じる場面もあったという。ただ首相は渡辺氏の造反に強気で臨んでいる。

C　判定例

必須箇所における判定例を以下に示す.

決定木，SVM 共に正しく分類された例

- 悠久の天山よりの雪解（ゆきげ）水連山を越えて彼方に冬の海「こうしたスケールの大きい風景の中に父を置くと、何とはなしにしっくりきます。あるいは自然の中で悠々と生きたかったのでしょうか」と堤さん。

決定木では正しく分類されたが SVM では正しく分類されなかった例

- 小堀は前半、相手の右アッパーに合わせて左フックを再三好打。だが、終盤は攻め手を欠いた。

決定木では正しく分類されなかったが SVM では正しく分類された例

- 斎藤通り魔殺人なんかは、そうかもしれません。また、自己愛を否定形でしか表せない人同士が連帯を希求するなら、共に滅びる方向しかないんじゃないでしょうか。

どちらも正しく分類されなかった例

- ０３年１２月の会見では「手術の結果は、かなりの程度、確実にがんは取りきることができたと思う、ということで、公務に復帰したころは、ＰＳＡの値も下降しており、回復のために明るい気持ちで散歩に励んでいました」と振り返った。また、宮内庁は０８年２月、「陛下が骨粗しょう症になる可能性がある」と発表。

D　要旨のはじめに（添削事例の本文）

文章を作成する能力というのは正確に物事を伝える上でとても重要である．その際，自然な文章にするためには適切な接続詞が必要である．我々は文章の自動添削システムを構築することを最終的な目標とし，その中でも特に判定が困難である，文間の接続詞（を分類した接続関係）の同定に関する研究を行った [1]．しかし，この研究では文間に接続詞が入るという前提で行ったため，文間に接続詞が必要ない場合への対処が問題となる．接続詞が必要ない文の同定を,「接続詞なし」というカテゴリーを追加して学習することも考えられるが，接続詞がない文間でも接続関係は存在する．また接続詞の用法として，接続詞がなければ意味が通らなくなるような接続詞と文章の構成として読みやすさを向上するための接続詞があり，それらを区別して推定することが求められる．そのため，接続詞が必要であるかどうかは接続関係の同定とは全く別の問題であると言える．そこで，どのような場合に接続詞が必要であるかを様々な観点から分析することで，任意の文書を与えた時に接続詞を必要とする箇所を自動的に判定する手法が必要であり，本研究ではその手法の開発を目的とする．

言語処理においては，接続詞は手掛かり語として使われることが主であり，接続関係を同定する研究 [1, 7] も少なく，ましてや接続詞がどこで必要となるかの研究は行われていない．しかし，言語学や国語教育などの分野においては，接続詞がどのような時に省略できるのかについて分析がいくつか行われている．市川 [9] はどのような接続詞が省略しやすいかを,「接続詞を挟む 2 文」と「ある文書内での接続詞」に対して調査を行っている．この調査では，文の構造や文脈の有無などに応じて接続詞が省略可能かどうかが決定されることが明らかになっている．

よって，本研究ではこれらの知見を考慮して，大量の実験データから接続詞が必要になる状況を分析し，接続詞の必要箇所を判定する手法を提案する．

E　接続関係の種類

接続関係の分類を省略せず表 25 に示す．また，表中の太字は用例利用型 [6] では複数の接続関係に出現していたが SVM を用いた研究 [1] では一意にしたものを表す．

表 25: 接続関係

接続関係	接続詞
並列	または，又は，あるいは，或いは，或は，もしくは，若しくは， それとも，ないし，乃至，つまり，すなわち，即ち，いっぽう，一方， かたや，および，及び，ないしは，ならびに，並びに，それから， それでいて，したがって，従って，よって
例示	たとえば，例えば，譬えば，たとへば，例へば，譬へば
因果	なので，だから，ですから，ゆえに，故に，ほんで，そこで， そやさかい，すると，だとすると，そうなると，そうすると， だとすれば，とすれば，だからこそ，それで，かくして，こうして， そうして，で，それだけに，したがって，従って，よって，それから， **そうしたら，したら，そしたら，では**
累加	また，又，亦，そして，それに，しかも，さらに，それから，そのうえ， そのうえに，それと，おまけに，ちなみに，因みに，なお，尚，なぜなら， ただし，但し，ただ，但，かつ，とどうじに，と同時に，あわせて，併せて， おなじく，同じく，それも，ましてや，それどころか，どころか， ついで，次いで，ならびに，まずは，つぎに，次に，そのうえで，だとすれば， とすれば，**ほな，ほなら，ほんなら，そういや，そういえば，そーいや，** それだけに，だって，というのも，**じつは，実は，ほんとうは，本当は，** もっとも，尤も，ともすれば，そもそも，**じゃ，んじゃ，しかしながら，** 然しながら，然し乍ら，それだけに，つまるところ，**そうですが**
転換	それでは，ところで，さて，それじゃ，じゃあ，ふんじゃ，ほんじゃ， それにしても，ともあれ，さあ，ならば，なら，それなら
逆説	しかし，然し，でも，ところが，だが，ですけれど， けれども，けれど，ですが，それでも，じゃが，だけど，されど，が， けど，だけれども，だからといって，さりとて，なのに，それなのに， にもかかわらず，さもなければ，でないと，いな，いや，否，いえ， そもそも，だからといって，はんめん，反面，逆に，でなければ， **てゆーか，てか，ってか**